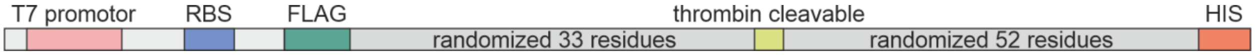
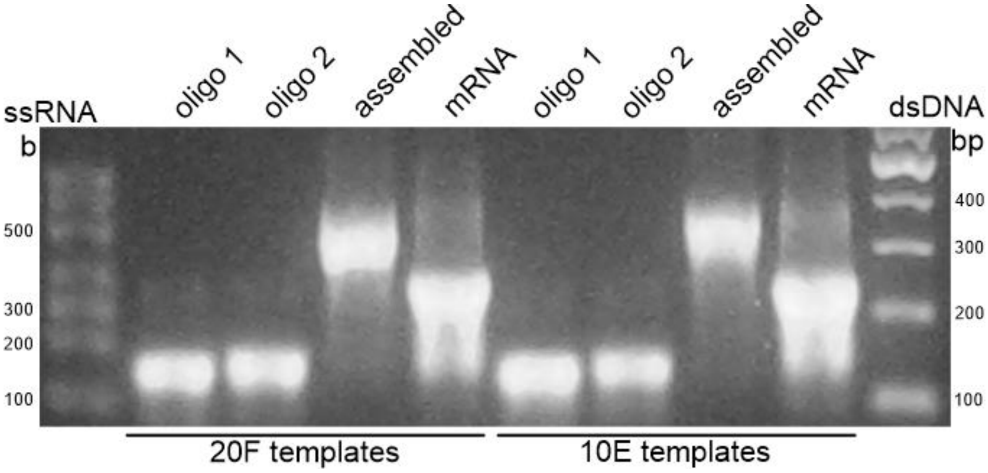
Structured proteins are abundant in unevolved sequence space

Vyacheslav Tretyachenko[1,2], Jiří Vymětal[3], Tereza Neuwirthová[1#], Jiří Vondrášek[3], Kosuke Fujishima[4,5] and Klára Hlouchová*[1,3]

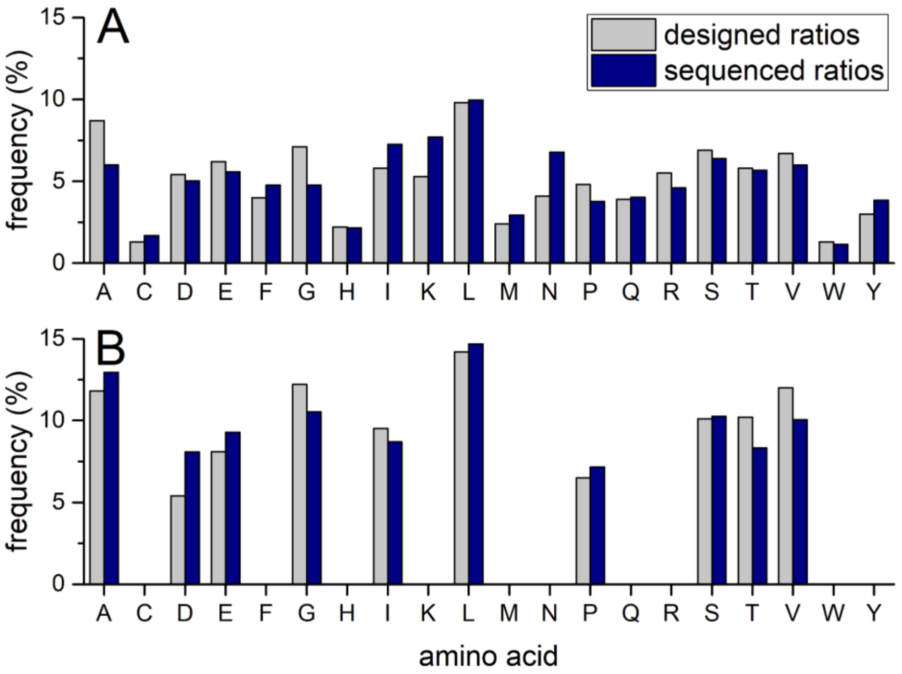
Supplementary Information



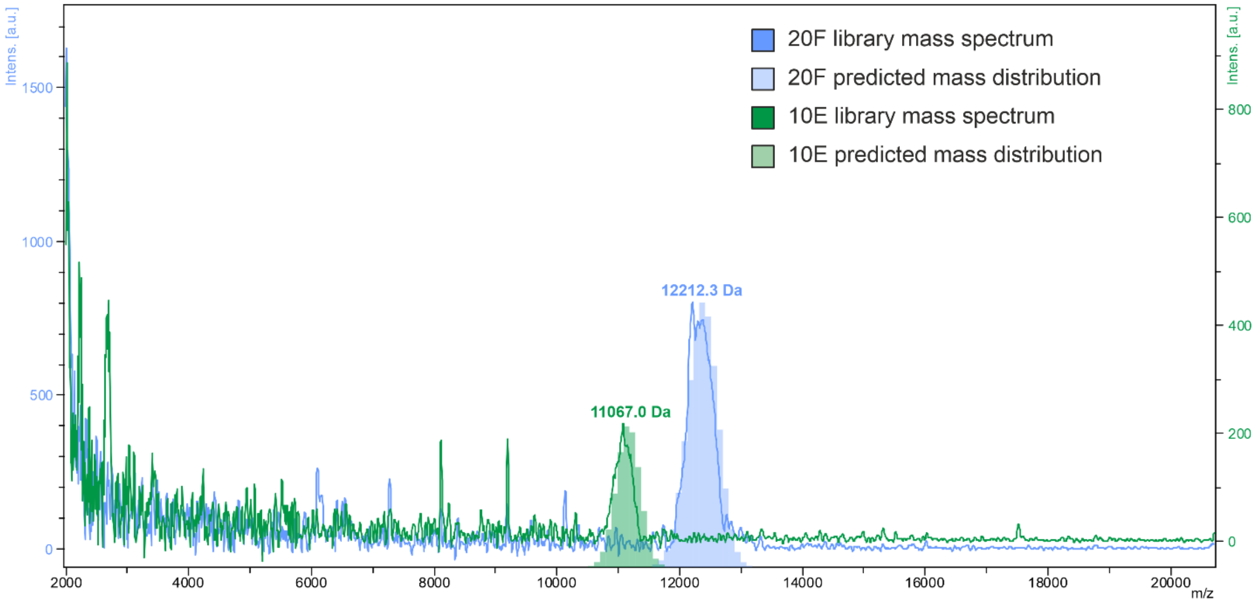
Supplementary Figure S1. The general scheme of the library expression cassette. Construct is provided with necessary sequences for in vitro transcription and translation, FLAG/HIS affinity purification tags on N/C ends of proteins and the thrombin cleavage site in the middle of the protein coding sequence



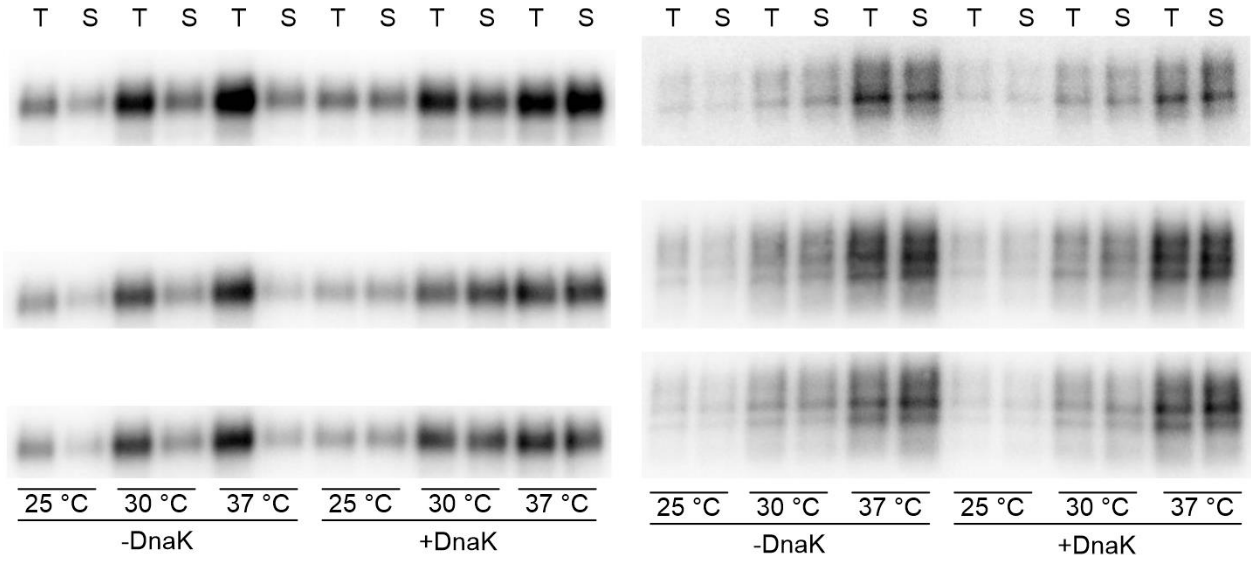
Supplementary Figure S2. Agarose gel representing degenerate ssDNA and assembled dsDNA library templates (A) and transcribed mRNA templates (B)



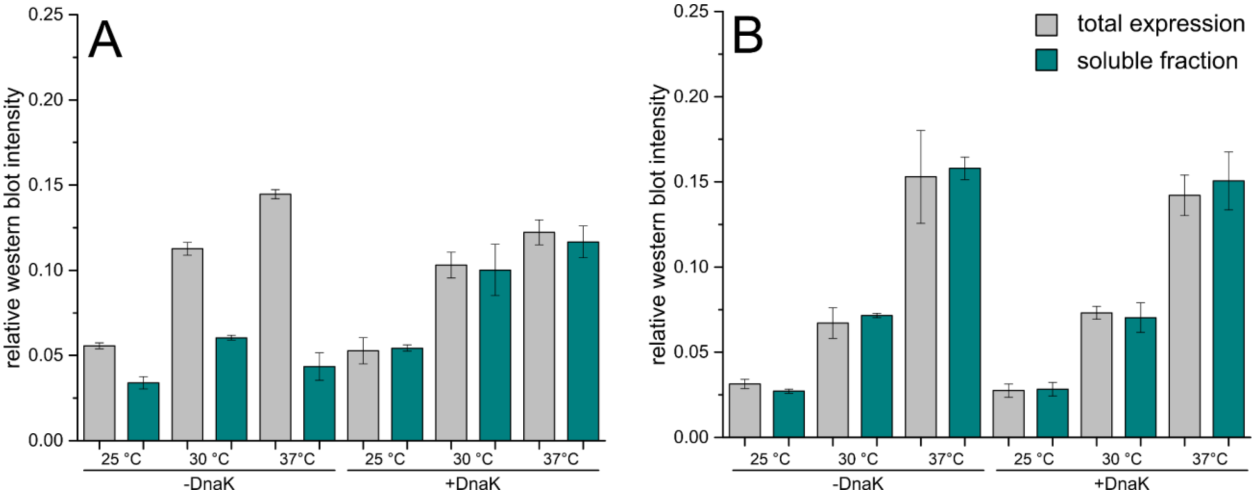
Supplementary Figure S3. Comparison of designed (grey) and experimental (blue) amino acid ratios in full (A) and early (B) alphabet libraries. Experimental amino acid distributions were calculated from the sequenced libraries DNA templates



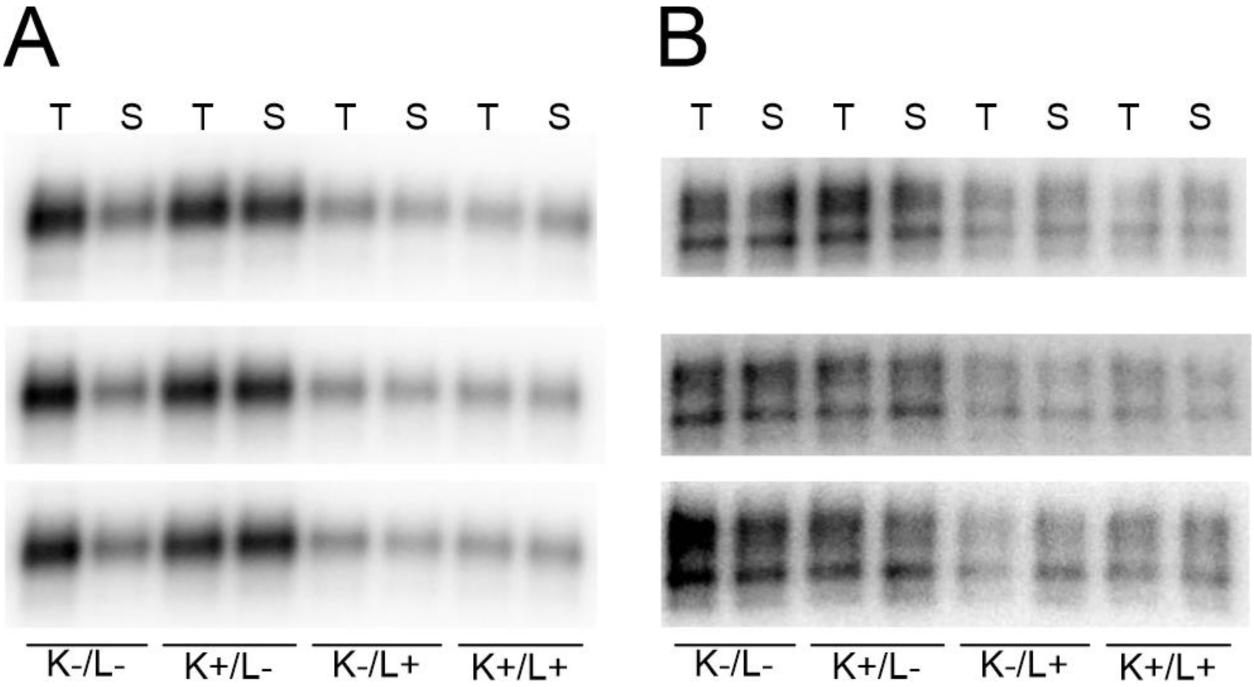
Supplementary Figure S4. Mass spectrometric analysis of purified full (blue) and early (green) alphabet libraries with their corresponding molecular weight distributions calculated in silico from the sequenced DNA templates



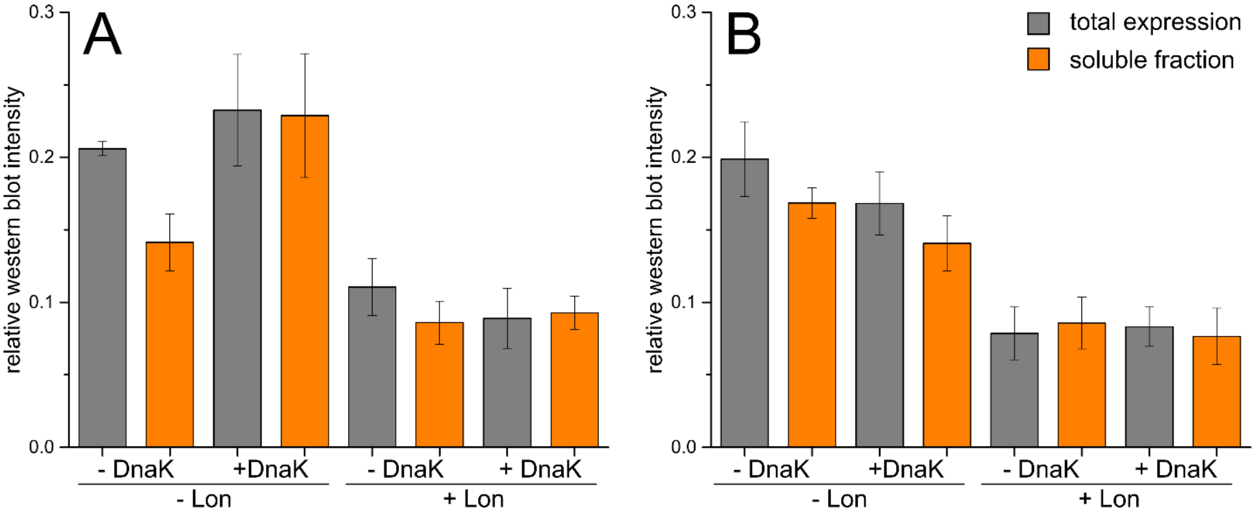
Supplementary Figure S5. Western blot signals used for temperature/chaperone dependent solubility analysis of full (A) and early (B) alphabet libraries. Total expressions (T) and soluble fractions (S) were analyzed



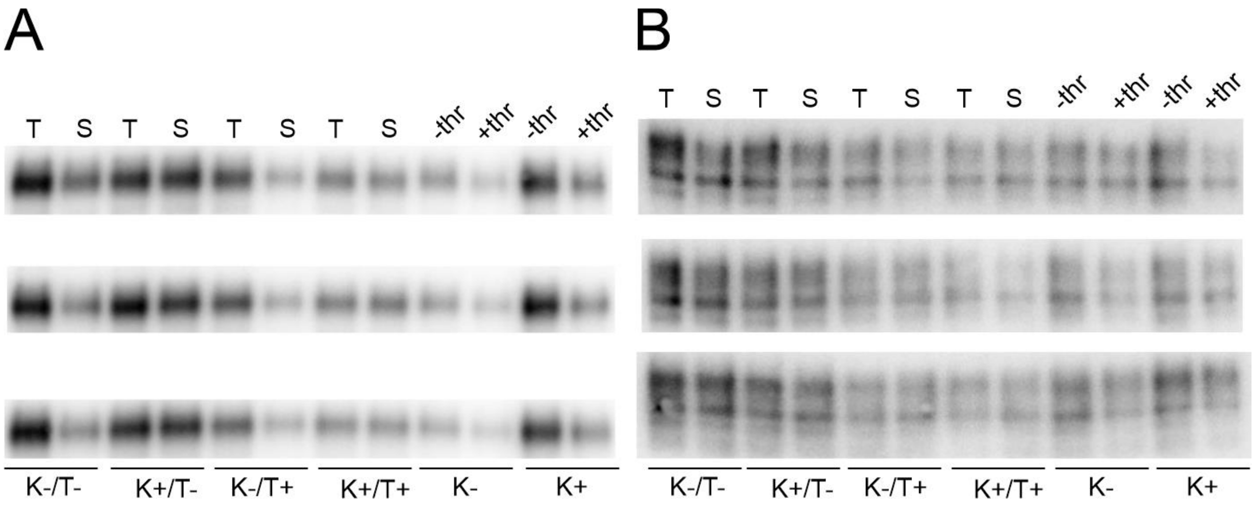
Supplementary Figure S6. Summary of triplicate western blot signal quantification of temperature/chaperone dependent solubility analysis of full (A) and early (B) alphabet libraries. Total expressions (grey) and soluble fractions (green) were analyzed



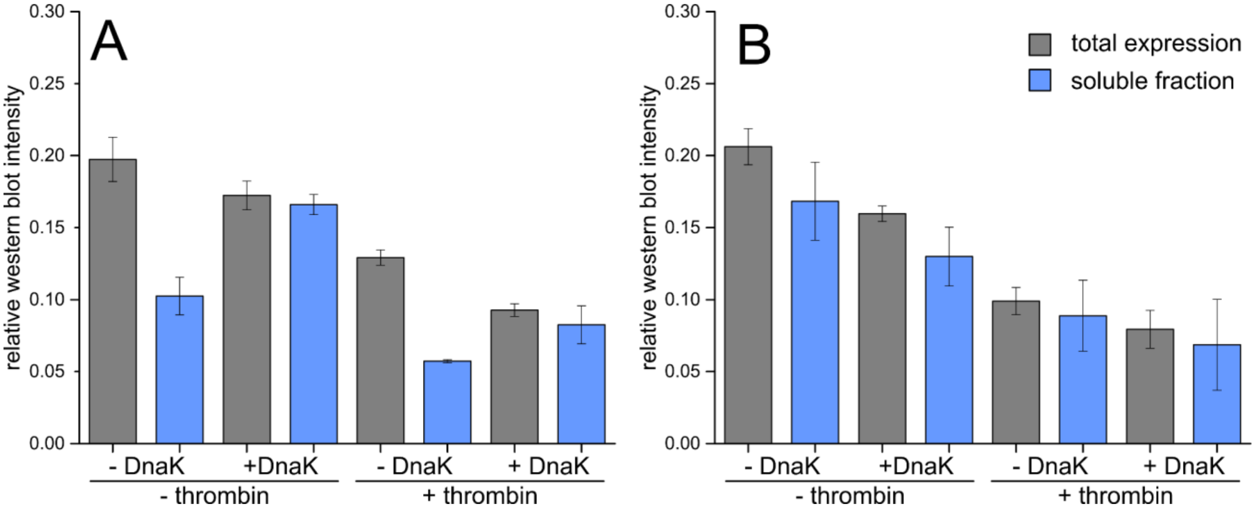
Supplementary Figure S7. Western blot signals used for Lon protease digestion/solubility analysis (K-/L-, K+/L-, K-/L+, K+/L+) and for the temperature dependent aggregation assay (42 °C) of full (A) and early (B) alphabet libraries. Libraries were expressed either in absence/presence of Lon protease (L-/L+) and absence/presence of DnaK/DnaJ/GrpE chaperone system (K-/K+). Total expressions (T) and soluble fractions (S) were analyzed. Only soluble fractions of chaperone absent (-K) or chaperone (+K) were analyzed after 42 °C heat shock treatment



Supplementary *Figure S8. Summary of triplicate western blot signal quantification of Lon protease digestion/solubility analysis of full (A) and early (B) alphabet libraries. Total expressions (grey) and soluble fractions (orange) were analyzed*



Supplementary Figure S9. Western blot signals used for thrombin protease digestion/solubility analysis (K-/T-, K+/T-, K-/T+, K+/T+) of full (A) and early (B) alphabet libraries. Libraries were expressed either in absence/presence of the DnaK/DnaJ/GrpE chaperone system (K-/K+) and treated/untreated (T+/T-) by thrombin protease. Total expressions (T) and soluble fractions (S) were analyzed.



Supplementary Figure S10. Summary of triplicate western blot signal quantification of thrombin protease digestion/solubility analysis of full (A) and early (B) alphabet libraries. Total expressions (grey) and soluble fractions (blue) were analyzed

Sequences

```
>20F_full

CTGTAATACGACTCACTATAGGGACACCAATAGAGAAAGAGGAGAAATACTAGATGGATTA
TAAAGATGATGATGATAAGKHYKHYRMKDGKWTNNYARRSSHRVDGRMKRRSKHYMWMD
GKNYAGRRRMKNYAVDGRMKNYASMWSMWMWMKHYKHYVDGVDGKHYNYAMWMNYAG
CGTTAGTCCCGCGTGGGAGCNYASMWRMKKHYKHYVDGNYARMKNYAKHYVDGRRSVDG
KGSHHYNYAKHYRRSWTNVDGHHYRRSHHYHHYRMKNYANYAHHYSHRVDGHHYKHYDGK
VDGMWMWTNHHYRMKVDGKGSNYAVDGMWMNYAVDGRRSRRSSHRVDGRRSSMWRMK
CACCACCACCACCACCACTAA
```

```
>10E_full

CTGTAATACGACTCACTATAGGGACACCAATAGAGAAAGAGGAGAAATACTAGATGGATTA
TAAAGATGATGATGATAAGGNTHYAGNGHYAGNGKCWGNGGNGHYAHYAHYAHYAHYAGN
THYAHYAHYAHYAHYAKCWGNGHYAGNTGAHHYAGNGHYAGNGHYAHYAHYAGNGGCGTT
AGTCCCGCGTGGGAGCGNTHYAGNGGNTGNGKCWGNGRBTHYAHYAGNGRBTHYAHYAG
NTHYAGNGHYAGNTHYAGNGRBTGNGGNGGNGRBTHYAHYAGAHHYAGNTGAHHYAHYAG
AHHYARBTGAHGNGHYAHYAGNTRBTHYAGNGGAHHYAGNGGMDGNTHYAGNGCACCAC
CACCACCACCACTAA
```

Supplementary Table S1

Amino acid ratios of constructed libraries 20F and 10E

Amino acid	ratio	Amino acid	ratio
(library 20F)		(library 10E)	
F	0.04	F	0
L	0.1	L	0.142
I	0.06	I	0.095
M	0.02	M	0
V	0.07	V	0.12
P	0.05	P	0.065
A	0.09	A	0.118
W	0.01	W	0
G	0.07	G	0.122
S	0.07	S	0.101
T	0.06	T	0.102
Y	0.03	Y	0
Q	0.04	Q	0
N	0.04	N	0
C	0	C	0
D	0.05	D	0.054
E	0.06	E	0.081
H	0.02	H	0
K	0.05	K	0
R	0.06	R	0

Supplementary Table S2

Statistics obtained by amino acid composition analysis of sequenced library constructs

amino acid	designed 10E	sequenced 10E	squared error	designed 20F	sequenced 20F	squared error
A	0.118	0.11138089	4.38126E-05	0.09	0.060045619	0.000897265
C	0	0.000126502	1.60028E-08	0	0.016746927	0.00028046
D	0.054	0.078464399	0.000598507	0.05	0.050126524	1.60083E-08
E	0.081	0.107400275	0.000696975	0.06	0.055882059	1.69574E-05
F	0	0.000180281	3.25013E-08	0.04	0.04771278	5.9487E-05
G	0.122	0.084749913	0.001387569	0.07	0.047688509	0.000497803
H	0	0.000134772	1.81636E-08	0.02	0.021467382	2.15321E-06
I	0.095	0.105720657	0.000114932	0.06	0.072494652	0.000156116
K	0	0.00033089	1.09488E-07	0.05	0.076937757	0.000725643
L	0.142	0.163457062	0.000460406	0.1	0.099681555	1.01407E-07
M	0	0.000206983	4.28421E-08	0.02	0.029316427	8.67958E-05
N	0	0.000206426	4.26117E-08	0.04	0.067735012	0.000769231
P	0.065	0.053218191	0.000138811	0.05	0.03771341	0.00015096
Q	0	0.001496622	2.23988E-06	0.04	0.040165308	2.73269E-08
R	0	0.000763828	5.83433E-07	0.06	0.045940967	0.000197656
S	0.101	0.100585852	1.71519E-07	0.07	0.063834005	3.80195E-05
T	0.102	0.083069409	0.000358367	0.06	0.056681099	1.10151E-05
V	0.12	0.108123364	0.000141054	0.07	0.059935965	0.000101285
W	0	0.000172568	2.97799E-08	0.01	0.011443048	2.08239E-06
Y	0	0.000210885	4.44727E-08	0.03	0.03845086	7.1417E-05
error			**0.062799398**			**0.063753359**

Supplementary Table S3

Statistics obtained from high throughput sequencing analysis

	total reads	unique reads	max. multiplicity	correct expressable constructs
F	1768293 (100 %)	1702198 (96 %)	4	1513985 (85 %)
E	1541180 (100 %)	1448384 (94 %)	5	1304780 (85 %)

Supplementary Table S4

Western blot signal intensities of 20F library solubility analysis

	1st	2nd	3rd	avg	SD	1st	2nd	3rd	avg	SD
F-25T	1373.415	781.368	975.018	1043.267	301.8665	0.054732	0.054593	0.057625	0.05565	0.001712
F-25S	936.668	492.89	511.651	647.0697	250.9749	0.037327	0.034437	0.030239	0.034001	0.003564
F-30T	2717.955	1636.547	1950.536	2101.679	556.3219	0.108313	0.114343	0.115279	0.112645	0.003781
F-30S	1549.558	863.842	997.947	1137.116	363.4247	0.061751	0.060355	0.05898	0.060362	0.001386
F-37T	3653.668	2025.941	2481.881	2720.497	839.6886	0.145602	0.141549	0.146682	0.144611	0.002706
F-37S	1321.08	534.543	688.156	847.9263	416.8993	0.052646	0.037348	0.040671	0.043555	0.008047
F+25T	1534.369	660.144	863.845	1019.453	457.4141	0.061146	0.046123	0.051054	0.052774	0.007658
F+25S	1407.286	749.519	923.859	1026.888	340.772	0.056082	0.052368	0.054601	0.05435	0.00187
F+30T	2586.142	1365.636	1871.581	1941.12	613.2173	0.10306	0.095415	0.110612	0.103029	0.007599
F+30S	2125.129	1640.581	1712.689	1826.133	261.4361	0.084688	0.114625	0.101222	0.100178	0.014995
F+37T	2855.814	1798.992	2150.483	2268.43	538.193	0.113807	0.125693	0.127096	0.122199	0.007301
F+37S	3032.406	1762.601	1792.534	2195.847	724.6359	0.120844	0.12315	0.105941	0.116645	0.009342
sum	25093.49	14312.6	16920.18	18775.42						

Western blot signal intensities of 10E library solubility analysis

E-25T	43.952	150.68	286.532	160.388	121.581	0.029971	0.02968	0.034564	0.031405	0.00274
E-25S	38.384	135.433	235.79	136.5357	98.70762	0.026174	0.026676	0.028443	0.027098	0.001192
E-30T	83.264	366.153	600.581	349.9993	259.0365	0.056778	0.072122	0.072447	0.067115	0.008954
E-30S	104.178	358.737	605.689	356.2013	250.7651	0.071039	0.070661	0.073063	0.071588	0.001291
E-37T	266.963	650.71	1231.772	716.4817	485.7556	0.182043	0.128171	0.148586	0.152933	0.027198
E-37S	222.918	837.598	1299.544	786.6867	540.1156	0.152009	0.164983	0.156761	0.157918	0.006564
E+25T	46.96	127.136	211.227	128.441	82.14128	0.032022	0.025042	0.02548	0.027515	0.00391
E+25S	42.889	121.517	262.612	142.3393	111.3316	0.029246	0.023935	0.031678	0.028287	0.00396
E+30T	112.807	370.676	576.208	353.2303	232.1926	0.076924	0.073013	0.069507	0.073148	0.00371
E+30S	104.569	310.392	650.473	355.1447	275.6899	0.071306	0.061138	0.078465	0.070303	0.008707
E+37T	191.77	784.223	1170.372	715.455	492.912	0.130769	0.154469	0.141179	0.142139	0.011879
E+37S	207.827	863.628	1159.168	743.541	486.9067	0.141718	0.17011	0.139828	0.150552	0.016964
sum	1466.481	5076.883	8289.968	4944.444						

F/E	library 20F or 10E
sign +/-	DnaK absent or present
T/S	total or soluble fractions

Supplementary Table S5

LON proteolysis/solubility analysis of 20F library

	1st	2nd	3rd	1st	2nd	3rd	avg	SD	avg corr
K-/L- T	2983.742	2884.68	3298.423	0.205459	0.201272	0.210875	0.205869	0.004814	0.205869
K-/L- S	1730.409	1648.647	1727.763	0.119155	0.115031	0.110459	0.114882	0.00435	0.114882
K+/L- T	2782.103	2635.104	2970.071	0.191574	0.183858	0.189883	0.188438	0.004055	0.188438
K+/L- S	2607.984	2762.893	2850.042	0.179584	0.192775	0.182209	0.184856	0.006982	0.184856
K-/L+ T	1293.049	1273.935	1410.853	0.089039	0.088886	0.090199	0.089374	0.000718	0.089374
K-/L+ S	1010.515	993.633	1092.843	0.069583	0.069329	0.069868	0.069593	0.00027	0.069593
K+/L+ T	951.59	1064.119	1164.28	0.065526	0.074247	0.074435	0.071402	0.00509	0.071402
K+/L+ S	1162.95	1069.23	1127.339	0.08008	0.074603	0.072073	0.075585	0.004093	0.071402
sum	14522.34	14332.241	15641.61						

Ratios (0-1) in chaperone absent condition

soluble/un	0.338047
soluble/de	0.219987
agg-prone/	0.345879
agg-prone/	0.096087

Ratios (0-1) in chaperone present condition		correction for higher solubility in soluble fraction
soluble/un	0.401115	0.378917
soluble/de	0.579874	0.602072
agg-prone/	0.041209	0.019011
agg-prone/	-0.0222	

K+/-	DnaK present/absent
L+/-	Lon present/absent
T/S	total/soluble fraction

Supplementary Table S5

LON proteolysis/solubility analysis of 10E library

	1st	2nd	3rd	1st	2nd	3rd	avg	SD	avg corr
K-/L- T	417.711	183.233	288.472	0.17375	0.196925	0.225233	0.198636	0.025784	0.198636
K-/L- S	390.431	168.136	208.236	0.162403	0.1807	0.162586	0.168563	0.010511	0.168563
K+/L- T	462.119	151.571	191.458	0.192222	0.162897	0.149486	0.168202	0.021856	0.168202
K+/L- S	340.848	147.924	155.268	0.141779	0.158978	0.12123	0.140662	0.018899	0.140662
K-/L+ T	211.682	83.808	73.787	0.088051	0.090071	0.057611	0.078578	0.018185	0.078578
K-/L+ S	239.674	60.808	117.547	0.099694	0.065352	0.091778	0.085608	0.017983	0.078578
K+/L+ T	164.161	80.357	121.404	0.068284	0.086362	0.09479	0.083145	0.013542	0.083145
K+/L+ S	177.46	54.632	124.601	0.073816	0.058714	0.097286	0.076605	0.019436	0.076605
sum	2404.086	930.469	1280.773						

Ratios (0-1) in chaperone absent condition		correction for higher solubility in soluble fraction
soluble/un	0.43098	0.395586
soluble/de	0.417623	0.453017
agg-prone/	0.186792	0.151397
agg-prone/	-0.03539	

Ratios (0-1) in chaperone present condition	
soluble/un	0.455437
soluble/de	0.380832
agg-prone/	0.12485
agg-prone/	0.038881

Supplementary Table S6

LON proteolysis/solubility analysis of 20F library

	1st	2nd	3rd	1st	2nd	3rd	avg	SD
K-/T- T	3000.18	3819.807	2691.214	0.1924924	0.2146316	0.1850885	0.1974042	0.0153718
K-/T- S	1830.98	1667.294	1401.612	0.1174764	0.0936838	0.096396	0.1025187	0.0130245
K+/T- T	2505.91	3174.4	2588.193	0.1607801	0.1783667	0.1780032	0.1723834	0.0100503
K+/T- S	2566.43	3090.356	2323.44	0.1646629	0.1736444	0.1597948	0.166034	0.0070258
K-/T+ T	2074.51	2189.923	1908.886	0.1331011	0.1230498	0.1312838	0.1291449	0.0053561
K-/T+ S	895.047	1039.228	816.991	0.0574265	0.0583933	0.0561886	0.0573361	0.0011051
K+/T+ T	1458.57	1561.697	1403.876	0.0935819	0.0877504	0.0965517	0.092628	0.0044775
K+/T+ S	1254.34	1254.338	1405.933	0.0804787	0.0704801	0.0966932	0.0825507	0.0132288
sum	15586	17797.043	14540.145					

Ratios (0-1) in chaperone absent condition

soluble/undegradable	0.2904505
soluble/degradable	0.2288838
agg-prone/degradable	0.1169005
agg-prone/undegradable	0.3637653

Ratios (0-1) in chaperone present condition

soluble/undegradable	0.4788784
soluble/degradable	0.4842889
agg-prone/degradable	-0.021626
agg-prone/undegradable	0.0584589

Western blot signal intensities used in thermostability assay 20F

	1st	2nd	3rd	1st	2nd	3rd	avg	SD
K-/T- T	3000.2	3819.807	2691.214	0.272796	0.289389	0.254646	0.272277	0.014188
K-/T- HS	1127.8	1058.334	966.027	0.102543	0.080179	0.091407	0.091377	0.00913
K-/T+ HS	631.25	670.65	589.314	0.057398	0.050808	0.055762	0.054656	0.002801
K+/T- T	2505.9	3174.4	2588.193	0.227854	0.240493	0.244898	0.237748	0.007224
K+/T- HS	2446.7	2972.882	2531.673	0.222472	0.225226	0.23955	0.229082	0.007487
K+/T+ HS	1286.1	1503.502	1202.02	0.116936	0.113905	0.113737	0.114859	0.00147
sum	10998	13199.575	10568.44					

Ratios (0-1) of total library expresion after heat shock treatment - chaperone absent

Soluble after heatshock	0.335601
Uncleaved after heatshock	0.200737

Ratios (0-1) of total library expresion after heat shock treatment - chaperone present

Soluble after heatshock	0.96355
Uncleaved after heatshock	0.483113

K+/-	DnaK present/absent
T+/-	thrombin present/absent
T/S	total/soluble fraction

Supplementary Table S6

LON proteolysis/solubility analysis of 10E library

	1st	3rd	5th grad	1st	3rd	5th grad	avg	SD
K-/T- T	344.386	489.768	438.41	0.218775	0.201046	0.1987	0.206174	0.010976
K-/T- S	226.658	444.347	394.329	0.143987	0.182401	0.178722	0.16837	0.021196
K+/T- T	271.972	402.14	311.879	0.172773	0.165076	0.141353	0.159734	0.016377
K+/T- S	175.465	342.11	304.136	0.111466	0.140434	0.137844	0.129915	0.016029
K-/T+ T	173.464	235.932	198.159	0.110195	0.096848	0.089812	0.098952	0.010353
K-/T+ S	105.589	248.516	214.716	0.067077	0.102014	0.097316	0.088802	0.018961
K+/T+ T	138.014	167.85	180.114	0.087675	0.068901	0.081633	0.079403	0.009583
K+/T+ S	138.606	105.432	164.643	0.088051	0.043279	0.074621	0.06865	0.022975
sum	1574.154	2436.095	2206.386					

Ratios (0-1) in chaperone absent condition

soluble/undegradable	0.430714
soluble/degradable	0.385926
agg-prone/degradable	0.134132
agg-prone/undegradable	0.049228

Ratios (0-1) in chaperone present condition

soluble/undegradable	0.42978
soluble/degradable	0.383538
agg-prone/degradable	0.119366
agg-prone/undegradable	0.067316

Western blot signal intensities used in thermostability assay 10E

	1st	3rd	5th grad	1st	3rd	5th grad	avg	SD
K-/T- T	344.386	489.768	438.41	0.27906	0.266954	0.244005	0.26334	0.014537
K-/T- HS	177.773	312.79	308.92	0.144051	0.17049	0.171935	0.162159	0.012818
K-/T+ HS	134.808	149.365	194.615	0.109236	0.081413	0.108317	0.099655	0.012905
K+/T- T	271.972	402.14	311.879	0.220382	0.219192	0.173582	0.204385	0.021787
K+/T- HS	224.577	313.302	328.411	0.181977	0.170769	0.182783	0.17851	0.005483
K+/T+ HS	80.578	167.285	214.49	0.065293	0.091181	0.119378	0.091951	0.022087
sum	1234.094	1834.65	1796.725					

Ratios (0-1) of total library expresion after heat shock treatment - chaperone absent

Soluble after heatshock	0.615778
Uncleaved after heatshock	0.378429

Ratios (0-1) of total library expresion after heat shock treatment - chaperone present

Soluble after heatshock	0.873399
Uncleaved after heatshock	0.44989